Physics and evolution of thermophilic adaptation

Igor N. Berezovsky and Eugene I. Shakhnovich*

Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, MA 02138

Edited by Alan R. Fersht, University of Cambridge, Cambridge, United Kingdom, and approved July 19, 2005 (received for review May 10, 2005)

Analysis of structures and sequences of several hyperthermostable proteins from various sources reveals two major physical mechanisms of their thermostabilization. The first mechanism is ''structure-based,'' whereby some hyperthermostable proteins are significantly more compact than their mesophilic homologues, while no particular interaction type appears to cause stabilization; rather, a sheer number of interactions is responsible for thermostability. Other hyperthermostable proteins employ an alternative, ''sequence-based'' mechanism of their thermal stabilization. They do not show pronounced structural differences from mesophilic homologues. Rather, a small number of apparently strong interactions is responsible for high thermal stability of these proteins. High-throughput comparative analysis of structures and complete genomes of several hyperthermophilic archaea and bacteria revealed that organisms develop diverse strategies of thermophilic adaptation by using, to a varying degree, two fundamental physical mechanisms of thermostability. The choice of a particular strategy depends on the evolutionary history of an organism. Proteins from organisms that originated in an extreme environment, such as hyperthermophilic archaea (*Pyrococcus furiosus*), are significantly more compact and more hydrophobic than their mesophilic counterparts. Alternatively, organisms that evolved as mesophiles but later recolonized a hot environment (*Thermotoga maritima*) relied in their evolutionary strategy of thermophilic adaptation on ''sequence-based'' mechanism of thermostability. We propose an evolutionary explanation of these differences based on physical concepts of protein designability.

thermostability | structure/sequence | molecular evolution | molecular packing | genomes/proteomes

The importance of various factors contributing to protein thermostability remains a subject of intense study (1). The most frequently reported trends include increased van der Waals interactions (2), higher core hydrophobicity (3), additional networks of hydrogen bonds (1), enhanced secondary structure propensity (4), ionic interactions (5), increased packing density (6), and decreased length of surface loops (7). It was shown recently that proteins use various combinations of these mechanisms. However, no general physical mechanism for increased thermostability was found. The diversity of the "recipes" for thermostability immediately raises two important questions: (*i*) What are possible physical mechanisms to increase thermostability of proteins, and (*ii*) how did evolution use possible physical mechanisms of thermal stabilization to develop strategies of adaptation to high temperature and other possible demands of the environment?

In this work, we first analyze in great detail several proteins from various hyperthermophilic organisms and show that some of them draw their thermostability from structural factors such as increased compactness. Furthermore, direct analysis of interactions as well as sequence comparison with mesophilic orthologues indicate that no specific forces apparently dominate interaction patterns in such proteins. On the other hand, we also found hyperthermophilic proteins that are even less compact than their mesophilic homologues. Those proteins appear to be stabilized by specific interactions like additional salt bridges. In this case, the physical mechanism of stabilization appears to be more related to sequence adjustment than to structural selection. Looking at sources of different proteins, we noticed a clear trend: Structure-stabilized proteins came mostly from archaea, whereas sequence-stabilized proteins were mostly from bacteria. Although "evidence" based on few proteins is anecdotal at best, it motivated us to carry out full high-throughput comparative structural and sequence analysis of several genomes and proteomes of hyperthermophilic organisms. This analysis pointed to a diversity of evolutionary strategies of thermophilic adaptation. We found that hyperthermophilic archaea used structure-based physical mechanisms to increase the thermostability of its proteins in the process of its thermophilic adaptation. Alternatively, some bacteria (such as *Thermotoga maritima*) used a "sequence-based" physical mechanism in their thermophilic adaptation. We attribute such differences to the vastly different phylogenetic histories of these organisms: The primordial habitat for archaea is believed to be a hot environment (8). When archaea evolved in such a habitat, its proteins were "*de novo*" designed in a hot environment that necessarily biased both structural repertoire (as explained in more detail below) and sequences that had to be found to fold and be stable in such structures. On the other hand, *T. maritima* is likely to have initially evolved as a mesophilic organism that later recolonized a hot environment (9). Its thermophilic adaptation required the enhancement of the thermostability of already existing proteins. Thus, our analysis reveals an intimate connection between the thermodynamics of protein structure, the evolution of thermophilic adaptation, and the phylogenetic history of an organism.

Materials and Methods

The set of proteins we have analyzed in this work consists of five groups (see *Supporting Text*, which is published as supporting information on the PNAS web site, for the listing).

X-ray data from the Protein Data Bank were supplemented with coordinates of H-atoms.

Unfolding simulations were performed by using an all-atom Gō model developed earlier (10). In the Gō interaction scheme, atoms that are neighbors in the native structure are assumed to have attractive interactions. Hence, the Gō model of interactions is structure-based. Every unfolding run consists of 2×10^6 steps. The move set contains one backbone move followed by one side-chain move.

van der Waals interactions were calculated for atoms belonging to residues separated by at least two residues along the polypeptide chain; only contact distances within 2.5–5.0 Å were considered for interactions.

High-throughput analyzes of the distributions of van der Waals contacts was performed on representative sets of major fold types [according to SCOP (http://scop.mrc-lmb.cam.ac.uk/scop) classification (11)] from *Aquifex aeolicus*, *Escherichia coli*, *T. maritima*, and *Pyrococcus furiosus/horikoshii/abyssi* (see listing of the fold in *Supporting Text*). A limited number of available folds (22, 37, and 42 for *A. aeolicus*, *P. furiosus/horikoshii/abyssi*, and *T. maritima*, respectively) is a caveat of the analysis. However, even these sets reveal a significant difference between mean values of distributions of number of contact per residue. We used normal distribution to estimate standard deviation and *P* values. Jack-knife tests were

This paper was submitted directly (Track II) to the PNAS office.

Freely available online through the PNAS open access option.

*To whom correspondence should be addressed. E-mail: eugene@belok.harvard.edu.

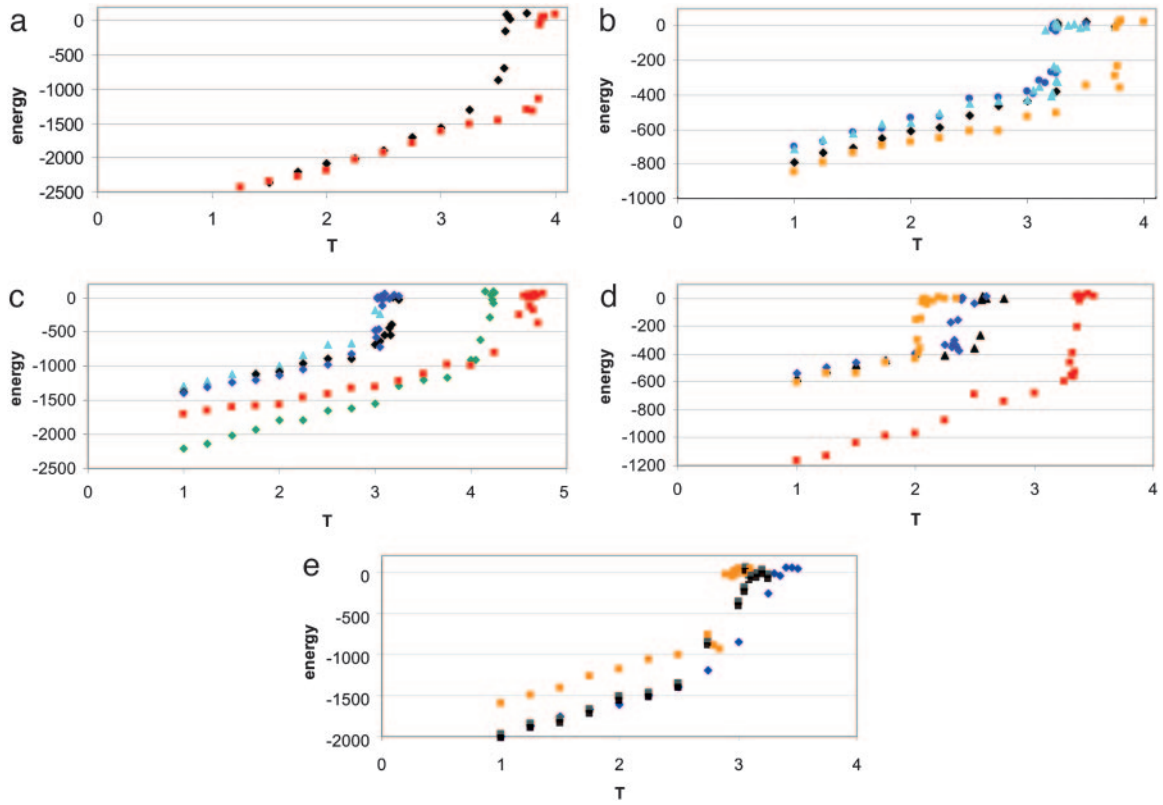

Fig. 1. The temperature-dependence of the energy of unfolding. Every simulation of unfolding started from the native structure and included 2×10^6 MC steps. The absolute temperature increment is 0.2, and 0.1 in the vicinity of transition temperature. In all plots, curves of the unfolding energy of mesophilic proteins are shown by black, blue, or cyan dots; thermophilic proteins, red dots; hyperthermophilic proteins, orange dots; halophilic protein, green dots. (*a*) Hydrolases, from *E. coli* (1INO, black rhombuses) and *Thermus thermophilus* (2PRD, red squares). (*b*) Rubredoxins, from *D. gigas* (1RDG, cyan triangles), *Clostridium pasteurianum* (5RXN, black rhombuses), *D. vulgaris* (8RXN, blue rhombuses), and *P. furiosus* (1CAA, orange squares). (*c*) 2Fe-2S ferredoxin, from *Spirulina platensis* (4FXC, cyan triangles), *Equisetum arvense* (1FRR, black rhombuses), *Anabaena PCC7120* (1FRD, blue rhombuses), *H. marismortui* (1DOI, green rhombuses), and *S. elongatus* (2CJN, red squares). (*d*) 4Fe-4S ferredoxin, from *C. acidiurici* (1FCA, black triangles), *Peptostreptococcus asaccharolyticus* (1DUR, blue rhombuses), *B. thermoproteolyticus* (1IQZ, red squares), and *T. maritima* (1VJW, orange squares). (e) Chemotaxis protein, from *E. coli* (3CHY, blue rhombuses), *Salmonella typhimurium* (2CHF, black squares), and *T. maritima* (1TMY, orange squares).

performed to exclude (*i*) possible effect of the same fold on the set and (*ii*) the influence of the size of the set.

Hydrogen bonds were determined according to criteria developed in ref. 12.

Sequence alignments were done by using the program MULTALIGN, developed in ref. 13. Sets of the best hits of *A. aeolicus* and *T. maritima* to archaea (archaeal part of genome) were extracted according to listing in taxonomic distributions of the homolog TaxMap (www.ncbi.nlm.nih.gov/sutils/taxik.cgi?gi=133 and www.ncbi.nlm.nih.gov/sutils/taxik.cgi?gi=141 for *A. aeolicus* and *T. maritima*, respectively). The rest of the genomes were considered as bacterial parts.

We used binomial law to estimate the difference between the occurrences of groups of residues (see *Supporting Text* for explanations). The total number of amino acids residues in proteomes and their archaeal and bacterial parts are as follows: *P. furiosus*, 587152; *T. maritima*, 584965; *A. aeolicus*, 483137; bacterial part of *T. maritima*, 433914; bacterial part of *A. aeolicus*, 391211.

Designability has been treated within the framework of a residue–residue contact Hamiltonian (14). It defines the conformational energy of a polypeptide chain to be the sum of the pairwise interaction energies of all of the amino acid pairs whose alpha carbons are separated by a distance of less than ≈7.5 Å.

Results

Unfolding Simulations with the Gō Model. First, we evaluated the stability of each of the proteins using an unfolding procedure based on the Gō model (15). The Gō model is a basic physical model that captures essential physical interactions in native structure, interactions in unfolded state, and entropic factors. According to the Gō model, native interactions in the structure of a protein reflect mutually stabilizing effects of all or almost all types of interactions.

It was demonstrated (15) that Gō-like models that consider only native interactions give a satisfactory description of two-state folding processes of single-domain proteins. Thus, Gō-model simulations aim at revealing *structure-based* contributions to protein stability, which means that they treat all native contacts equally and are not able to detect stabilization due to a small number of especially strong but specific interactions (15).

Here, we used Monte Carlo unfolding simulations with the Gō model to analyze five groups of proteins, each of them containing representative(s) of mesophilic organisms and its homologues from (hyper)thermophilic species. Unfolding simulations for the studied groups of proteins reveal general trends of higher observed transition temperatures of unfolding for several (hyper)thermophilic proteins compared with their mesophilic counterparts. Fig. 1*a* shows the difference between hydrolases from thermophilic *Thermus thermophilus* and mesophilic *E. coli* toward higher stability of thermophilic protein. There is a pronounced difference between the unfolding temperatures of the rubredoxin from hyperthermophilic *P. furiosus* and rubredoxins from three mesophilic organisms (Fig. 1*b*). Three mesophilic 2Fe-2S ferredoxins (Protein Data Bank entries 4FXC, 1FRR, and 1FRD) show a narrow range of transition temperatures, whereas the thermophilic one (2CJN) from cya-

BIOPHYSICS

Table 1. Factors possibly contributing to thermostability of analyzed proteins

Protein	vdW cnts/res	Hydrogen bnds/res	Secondary structure
Hydrolase			
1INO (175)	130	0.83	0.48
2PRD (174)	133 (2.3)	0.98 (18.1)	0.6 (25)
Rubredoxin			
1RDG (52)	103	0.77	0.40
5RXN (54)	98	0.72	0.39
8RXN (55)	96	0.76	0.4
1CAA (53)	112 (13.1)	0.85 (13.3)	0.62 (56.3)
Ferredoxin (2FE-2S)			
4FXC (98)	113	0.78	0.37
1FRR (95)	124	1.01	0.43
1FRD (98)	123	1.04	0.49
1DOI (128)	137 (14.2)	1.02 (8.1)	0.5 (16.3)
2CJN (97)	138 (15.0)	—	0.56 (30.2)
Ferredoxin (4FE-4S)			
1FCA (55)	96	0.71	0.22
1DUR (55)	82	0.67	0.4
1IQZ (81)	113 (27.0)	0.90 (30.4)	0.44 (41.9)
1VJW (59)	95 (6.7)	0.97 (40.6)	0.49 (58.1)
Chemotaxis protein			
3CHY (128)	135	1.28	0.58
2CHF (128)	136	1.3	0.58
1TMY (118)	131 (−3.3)	1.14 (−11.6)	0.7 (20.7)

van der Waals interactions (29), number of H-bonds (12, 30), and number of residues involved into elements of secondary structure in groups of proteins under consideration. vdW cnts/res, number of vdW contacts per residue; Hydrogen bnds/res, number of H-bonds per residue; Secondary structure, percentage of residues involved into the elements of secondary structure. Names of (hyper)thermophilic organisms in the second column are bolded italic. Numbers in brackets show difference between numbers of vdW contacts per residue, H-bonds per residue, and number of residues involved into secondary structure in mesophilic (averaged value was used if there are several mesophilic proteins in the group) and (hyper)thermophilic proteins, respectively.

nobacterium *Synechococcus elongatus* has a substantially higher temperature of unfolding (Fig. 1*c*). Analysis of 4Fe-4S ferredoxins from mesophilic and thermophilic organisms also reveals a significant difference in their transition temperatures (Fig. 1*d*), pointing to increased thermostability of thermophilic ferredoxin (1IQZ).

Proteins from hyperthermophilic *T. maritima*. Both 4Fe-4S ferredoxin (1VJW) and chemotaxis protein, CheY (1TMY), represent a striking exception from the general rule of higher simulation transition temperature for (hyper)thermostable proteins: They exhibit lower transition temperatures than their respective mesophilic counterparts (Fig. 1 *d* and *e*). Therefore, the Gō model discriminates between proteins from *T. maritima* and other proteins. This result apparently shows that mechanism of thermal stabilitization for ferredoxin and CheY protein from *T. maritima* may be *different* from those of other (hyper)thermostable proteins studied in our unfolding simulations.

Structural Analysis. According to the data in Table 1, hydrolase from the thermophilic bacteria has a higher total number of van der Waals contacts compared with its mesophilic counterpart. There are six α-helices in thermophilic protein and only three α-helices in the mesophilic one. Elements of secondary structure in thermostable hydrolase (2PRD) are rather extended in size, and the density of hydrogen bonds is also higher in a protein from the thermophilic organism (Table 1). Thus, according to all structural factors presented in Table 1, hydrolase from *Thermus thermophilus* is expected to be more stable compared with its mesophilic counterpart. This finding also agrees with experimental data (16) where the role of hydrophobic interactions in core region of thermophilic hydrolase was proven as a crucial factor of stabilization. Hyperthermophilic rubredoxin from the archaebacteria *P. furiosus* has a pronounced bias toward enhanced packing density compared with mesophilic proteins (Table 1). The higher density of packing in hyperthermophilic proteins is also reflected in the increased number of H-bonds per residue and in the involvement of 62% of residues into elements of secondary structure compared with 39–40% in mesophilic proteins. Van der Waals interactions and involvement of more residues into elements of secondary structure contribute to an increase of stability of thermophilic 2Fe-2S ferredoxin (2CJN; H-bonds cannot be obtained because of low-resolution NMR structure), in agreement with the conclusion made in experimental work (17). All major structural factors presented in Table 1 point to increased thermostability in thermophilic 4Fe-4S ferredoxin (1IQZ) and, thus, explain its higher transition temperatures in unfolding simulations compared with mesophilic analogues.

Proteins from *T. maritima* reveal a principally different distribution of major stabilizing interactions (Table 1). Analysis of the data for 4Fe-4S ferredoxin (1VJW) gives a substantially increased number of hydrogen bonds and involvement of almost half of the residues into secondary structure elements. At the same time, the compactness of the structure (Table 1) is practically the same as those in mesophilic protein. CheY protein (1TMY) has a lower density of van der Waals contacts and hydrogen bonds, and a slightly higher fraction of residues participating in secondary structure (see Table 1).

Sequence Analysis. Similarly to unfolding simulations and structural analysis, sequence alignments discriminate proteins from hyperthermostable *T. maritima* from other (hyper)thermostable proteins analyzed in this work. They have lower sequence identity with respective mesophilic proteins and show substantial redistribution or increased number of charged residues (see Fig. 3 and Table 5, which are published as supporting information on the PNAS web site). Contrary to *T. maritima*'s proteins, thermophilic hydrolase (2PRD, from *Thermus thermophilus*), ferredoxins (2CJN and 1IQZ, from *S. elongatus* and *Bacillus thermoproteolyticus*, respectively), and hyperthermophilic rubredoxin (1CAA, from *P. furiosus*) exhibit a high level of sequence identity with their mesophilic orthologues and demonstrate no significant substitutions into charged residues in their sequences.

Detecting Distinct Mechanisms of Thermostability in Individual Proteins. Both unfolding simulations (Fig. 1) and structural analysis (Table 1) show that the increased stability of thermophilic hydrolase (2PRD), ferredoxins (2CJN and 1IQZ), and hyperthermophilic rubredoxin (1CAA) from *P. furiosus* is provided by the majority of structural factors acting together on the background of increased compactness of their structures. We have checked for possible contribution of loop shortening to thermostabilization, because this mechanism was also suggested (7). We did not find significant deletions in the structures of (hyper)thermophilic proteins versus their mesophilic counterparts (see alignments of sequences in *Supporting Text*). On the contrary, in the single case of two-residue deletion in loop 6 in CheY protein from *T. maritima* (18), we found lower density of van der Walls interactions (Table 1). Thus, ferredoxin and CheY proteins from hyperthermophilic *T. maritima* do not reveal structural basis in their mechanisms of stability. Sequence analysis (Fig. 3 and Table 5), in turn, uncovers another possible mechanism of thermostability in *T. maritima*'s proteins. The stability of these proteins under extremely high temperatures is apparently provided by significant modifications of their sequences toward enrichment by charged residues (19, 20), which can be an effective sequence-based method of adaptation to extreme specific conditions.

Table 2. Comparative analyzes of the distributions of van der Waals contacts in representatives of the major fold types from *A. aeolicus, E. coli, P. abyssi/horikoshii/furiosis*, and *T. maritima*

Source	*Ec* (124 ± 5.4)	*Aa* (133 ± 7.3)	*Tm* (127 ± 4.2)	*Pr* (139 ± 5.1)
Ec		$1{\cdot}10^{-4}$	$4{\cdot}10^{-2}$	$<10^{-14}$
Aa			$6{\cdot}10^{-3}$	$6{\cdot}10^{-3}$
Tm				$6{\cdot}10^{-13}$

In parenthesis there are mean values approximated by normal distribution with three standard deviation intervals. *P* values show differences between mean values of the contact distribution for respective species. *Aa*, *Ec*, *Pr*, and *Tm* are *A. aeolicus* (22 folds), *E. coli* (34), *P. abyssi/horikoshii/furiosus* (37), and *T. maritima* (42), respectively.

Table 3. Comparison of the percentage of groups of amino acid residues in *T. maritima* and *A. aeolicus* proteomes compared with those in *P. furiosus* (boldface)

Source	Groups of amino acid residues		
	DEKR	ACFILMPVWY	HNQST
P. furiosus	**26.7**	**50.1**	**16.1**
T. maritima	27 (5)	48.7 (21)	17.4 (27)
A. aeolicus	28.3 (25)	48.7 (19)	16.2 (1.9)

Difference in σs estimated with binomial law is presented in parentheses. Groups of amino acid residues include charged (DEKR), hydrophobic (ACFILMPVWY), and hydrophilic (HNQST). G is not included in any of them.

Thus, all aspects of our analysis consistently distinguish individual proteins by mechanisms of gaining thermostability. In the first case, the structure-based Gō model detects the increase of transition temperature for (hyper)thermophilic proteins in unfolding simulations. Here, all stabilizing structural factors act in concert, pointing to enhanced compactness as the most probable original cause for higher stability. In the second case, on the contrary, we found a strong sequence bias that can explain dominating role of some of the stabilizing interactions, e.g., electrostatics (19, 20), but not of others. The high level of sequence variation compared with mesophilic orthologs and the significant bias toward charged residues in their sequences point to a key role of sequence selection in adaptation of *T. maritima* proteins (1TMY and 1VJW) to extreme conditions of the environment, in contrast to other (hyper)thermophilic proteins (1CAA, 1IQZ, 2CJN, and 2PRD) where structural bias is more pronounced.

From Physical Mechanism of Thermal Stabilization to Strategies of Thermophilic Adaptation of Organisms. The difference between physical mechanisms of stabilization is rather suggestive, showing that distinct physical mechanisms can be used *en route* of protein evolution: (*i*) on the basis of nonspecific compactness that increases the sheer number of interactions in protein structure, or (*ii*) through the sequence modification, using unique physical chemical features of amino acid residues. However, a conclusive evidence of sequestering of different physical mechanisms into distinct evolutionary strategies can be obtained only from massive comparison of protein structures and sequences from different species. We performed a twofold high-throughput analysis: (*i*) comparison of packing densities in proteins crystallized from *E. coli* (archetypal mesophilic bacteria), *A. aeolicus* and *T. maritima* [hyperthermophilic bacteria that recolonized hot environment (9, 21)], and *P. abyssi/horikoshii/furiosus* (archetypal archaea); and (*ii*) survey of amino acid compositions for respective complete genomes, which represent distant branches of the phylogenetic tree, archaea and bacteria. According to Table 2, proteins of archaeal *Pyrococcus* are most densely packed (139 contacts per residue), and respective *P* values distinguish its packing density from those of *E. coli*, *A. aeolicus*, and *T. maritima* folds. There are also other indications that, although *P. furiosus* and pair *A. aeolicus*/*T. maritima* are hyperthermophilic organisms (see Table 6, which is published as supporting information on the PNAS web site), they apparently developed different mechanisms of adaptation to hot environment. *A. aeolicus* and *T. maritima* have more charged residues than *P. furiosus*, whereas the latter has significantly elevated, compared with *A. aeolicus* and *T. maritima*, content of hydrophobic residues. Thus, both amino acid content and packing density show a difference between hyperthermophilic archaea (Pyrococcus) and bacteria (*A. aeolicus* and *T. maritima*). Increased packing density in archaea correlates with an increased contact density observed in several thermophilic proteomes (22) and higher contact density for the last universal common ancestor (LUCA) domains/folds (23). Together with results of unfolding simulations (Fig. 1*a*) and structural analysis (Table 1), it points to compactness as a key factor in the structure-based strategy of thermophilic adaptation in archaea. At the same time, lower packing density in proteins from *A. aeolicus* and *T. maritima* and a higher (compared with mesophilic *E. coli*, as well as with hyperthermophilic *P. furiosus*) content of charged residues suggest that these organisms follow mostly sequence-based stabilization, possibly, with a key role of ionic interactions (19, 20).

Comparison of archaeal versus bacterial parts (see *Materials and Methods* for definition and description of the comparison) of *A. aeolicus* and *T. maritima* detects, in addition, phylogenetic difference in their evolutionary history. The archaeal parts of bacterial genomes are the result of lateral (horizontal) gene transfer. Accordingly, the sequences that bacteria received from archaea upon recolonization are expected to preserve signals of archaeal mechanism of thermostability. In particular, a comparison of amino acid compositions of archaeal/bacterial parts of genomes should reflect phylogenetic history of bacterial genomes and phylogenetic distance to archaea. We found that there is a difference in the percentages of hydrophobic, hydrophilic, and charged (slight increase in bacterial part) residues between archaeal and bacterial parts of *A. aeolicus* (Table 4). However, the difference in amino acid composition between archaeal and bacterial parts of *T. maritima* is even more striking. It is of the same magnitude as one between hyperthermophilic archaea *P. furiosus* and bacteria *A. aeolicus* and *T. maritima* (Table 3). A plausible explanation for these observations comes from phylogenetic analysis. Indeed, *A. aeolicus* is a deeply branched hyperthermophilic bacteria, separated from the rest of bacteria kingdom at early stages of evolution and located closer to archaea (21). This fact explains the higher similarity of amino acid composition of archaeal and bacterial parts of *A. aeolicus*. *T. maritima*, on the contrary, recolonized a hot environment later; therefore, the striking difference in amino acid compositions between its archaeal and bacterial parts (Table 4) corroborates long evolutionary distance between *T. maritima* and archaea.

Discussion

Earlier studies of the mechanisms of protein thermostability resulted in the discovery of a variety of contributions to the effect

Table 4. Comparison of the percentage of groups of amino acid residues in archaeal (a, boldface) and bacterial (b) parts of *A. aeolicus* and *T. maritima*

Source	Groups of amino acid residues		
	DEKR	ACFILMPVWY	HNQST
***A. aeolicus* (a)**	**29.0**	**48.3**	**15.8**
A. aeolicus (b)	28.1 (12)	48.8 (6)	16.2 (7)
***T. maritima* (a)**	**25.4**	**50.1**	**17.4**
T. maritima (b)	27.4 (30)	48.4 (22)	17.2 (3.4)

For explanations see the legend to Fig. 3.

BIOPHYSICS

(2–7) and corresponding models on the basis of their combinations (1). However, the diversity of protein folds of thermostable proteins, the mechanisms of stability, and the evolutionary history of respective species raised questions about role of particular interactions or their combinations. The elusiveness of universal rules of thermostability stems from the long-standing tendency to contrast the role of different stabilizing interactions, e.g., hydrophobic versus ionic interactions. Furthermore, many researchers attributed a key role in stabilization under high temperatures exclusively to ionic interactions (4). If that would be true, then one would have to universally observe the prevalence of electrostatic stabilization in all thermostable proteins. However, this is not the case for several proteins studied here (see Fig. 1 and Table 1). High-throughput analysis on a proteomic level reinforces this observation (see Tables 2–4 and 6), showing an apparent key role of increased packing density in achieving the thermostability of proteins from hyperthermophilic archaea in contrast to a decrease of compactness coupled with sequence bias toward charged residues in *A. aeolicus* and *T. maritima*. Importantly, the percentage of charged residues in hyperthermophilic organisms is highly elevated. The increase of the number of charged residues in hyperthermophilic proteomes appears to be much greater than would have been necessary for stability purposes alone. Indeed, enhanced stability can be achieved by the addition of only a few ion pairs (4, 19, 20). This points to the possibility of alternative reasons (unrelated to protein stabilization) for this specific compositional bias toward charged residues, which should be thoroughly explored.

Discriminative Power of the Gō Model. Here, we demonstrated how simple all-atom simulations can be used to estimate the relative thermostability of proteins in the case of a structure-based mechanism of stabilization. We considered proteins from species with different growth temperature: mesophilic (growth temperature up to 60°C), thermophilic (up to 80°C), and hyperthermophilic (>80°C). By analogy with microcalorimetric experiments (24), where the transition temperature of unfolding is used as one of the parameters to evaluate protein thermostability, we compared the transition temperatures of unfolding obtained in simulations on the basis of the Gō model (15). It should be noted that the Gō model is a simple structure-based approach and, thus, reflects mostly the enthalpic contribution to free energy correlated with the compactness of the structure and opposing entropic factors arising from backbone and side-chain degrees of freedom. The model is neither supposed to predict transition temperature, nor to describe the dependence of hydrophobic or electrostatic interactions on temperature. Our aim here was to discriminate between robust vs. sequence-dependent physical mechanisms of thermostability, and we showed that the Gō model is a proper tool to achieve that end. We found that more dense proteins (hyperthermophilic rubredoxin from *P. furiosus*, thermophilic hydrolase from *Thermus thermophilus*, 2Fe-2S ferredoxin from *S. elongatus*, and 4Fe-4S ferredoxin from *B. thermoproteolyticus*) unfold at higher temperatures in Gō simulations. Failure of Gō model simulations to detect the higher unfolding temperature of certain thermophilic proteins indicates a possibility of an alternative mechanism of their specific stabilization, whereby protein sequences are selected in such a way to enhance only one or few types of interactions to adapt to very specific extreme conditions. In this case, sequence variation is responsible for the formation of specific stabilizing interactions, e.g., ion pairs (5), regardless of the details of the original structure, and this feature is not captured by the Gō model. Hyperthermophilic ferredoxin and chemotaxis protein from *T. maritima* exemplify this mechanism of stabilization (19, 20). Here, the obvious sequence bias couples with lack of nonspecific structure-based stabilization. Structural (Table 1) and sequence (Fig. 3 and Table 5) analysis further confirmed the existence of two physical mechanisms underlying thermostabilization: (*i*) increase of compactness so that all stabilizing interactions contribute to enhanced thermostability, and (*ii*) sequence-based formation of few strong interactions via sequence modification.

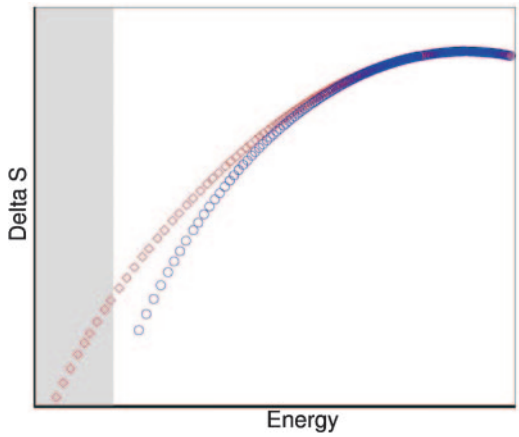

Fig. 2. Difference of sequence space entropy $S(E)$ from its maximum value as a function of energy. Sequence space entropy $S(E)$ represents the logarithm of the number of sequences that can fold into a given structure with a given energy E. Red diamonds show $S(E)$ for a more designable structure of high contact trace (or higher compactness in lowest order approximation), and blue circles correspond to a structure of low contact trace. A greater number of low-energy (thermostable) sequences can be ''accommodated'' by higher trace structures (gray shaded region), and, therefore, such structures can adopt a much larger number of foldable, highly thermostable sequences. The curves presented are for illustrative purposes only; detailed calculations for several specific models are presented in ref. 14.

Causal Relationship Between Physics of Mechanisms of Thermostability and Strategies of Organismal Adaptation. Up to this point, we discussed mechanisms of thermostability of individual proteins. What patterns emerge when we explore thermophilic adaptation on the organismal level? In other words, what is the causal relationship between mechanisms of thermostability of individual proteins and adaptation at the level of genomes/proteomes? Apparently, evolution sequestered distinct physical mechanisms for the developments of two major strategies, structure-based and sequence-based, according to the following possible scenario. The common belief that life started from hot conditions (8) implies two possible ways of evolutionary adaptation to hot environment: (*i*) organisms whose adaptation mechanisms should be developed "from scratch", i.e., simultaneously with discovery of new structures for their proteomes, whereas (*ii*) some organisms could have evolved as mesophiles but on later stages recolonized an extreme environment (9, 21) and, then, their already existing proteins should be changed. In the first scenario, thermostable proteins were designed *de novo*: selection of sequence and structure had to occur concomitantly. This process gives rise to evolutionary pressure on protein structures to make them more designable. Designability is a property of a protein structure that indicates how many sequences exist that fold into that structure at various levels of stability (14, 25).

Theoretical treatment of designability considers certain properties of contact matrix of a structure, C (14), as a major structural determinant of protein designability. Traces of powers of C reflect topological characteristics of the network of contacts within the structure and, as a consequence, determine the number of low-energy sequences that a fold can accommodate (14). In particular, in the lowest, second order in C, approximation, designability is predicted to correlate simply with compactness of a structure—number of contacts per residue (contact density) (14). Fig. 2 shows that higher trace, i.e., more compact, structures (red diamonds) can obviously accommodate more low-energy sequences (Fig. 2, gray shaded portion at left) than those of low contact trace, i.e., less

compact structures (blue circles). This finding suggests that more designable structures were more amenable to becoming thermostable proteins at the early stages of evolutionary selection, when structures and sequences were selected concomitantly: More designable structures had initial advantage because a greater number of sequences can fold into them with low energy, resulting in less severe sequence search requirements to make thermostable proteins having that structure. This fact, coupled with the earlier observation of higher contact density for last universal common ancestor (LUCA) domains (23), suggests that nature used higher designability in the creation of the first thermostable proteins of ancient species. Archaea proteins, rubredoxin from *P. furiosus* and 2Fe-2S ferredoxin from *Haloarcula marismortui*, exemplify this ancient mechanism of thermophilic adaptation, through selection of more compact (i.e., highly designable) structures (14). Finally, massive analysis of major folds reveals a statistically significant increase of packing density in archaeal Pyrococcus, compared with either mesophilic (*E. coli*) or hyperthermophilic (*T. maritima*) bacterial folds (Table 2). Thus, on the organismal level, the compactness of the ancient folds made it possible to adopt a great amount of different sequences and, as a consequence, to select those which are more stable. This structure-based mechanism was developed in the beginning of protein evolution and gave rise to the respective strategy of thermophilic adaptation (14, 22).

The second scenario is a modification of existing proteins of an organism in response to abruptly changed conditions of the environment. The fast and effective way of tuning of protein stability without redesign of the whole structure is to make sequence substitutions that would lead to formation of a "staple," a restricted set of specific interactions (e.g., ion bridges). This scenario gives rise to a sequence-based strategy of thermophilic adaptation. A good example of such strategy is *T. maritima* that recolonized a hot environment (9). A whole-genome similarity comparison demonstrates (9) that *T. maritima* has *only* 24% of genes that are most similar to archaea's. This similarity is a consequence of lateral (or horizontal) gene transfer (9, 21), which, as it was demonstrated earlier, points to specific biochemical and environmental adaptations (26). In this case, Archaea served as a source for lateral gene transfer on an organismal level of adaptation during recolonization (9), which was detected by comparison of the archaeal and bacterial parts of the *T. maritima* genome (Table 4). However, the mechanism of thermostabilization of the remaining, biggest part of its proteome should be developed, upon its colonization of hot environment, in *T. maritima* itself. In other words, when *T. maritima* recolonized a hot environment, the stability of the already existing proteins must be significantly improved. We showed here a crucial role of a sequence-based strategy to achieve thermostability in proteins from *T. maritima* versus the structure-based one in Archaea proteins (see *Results*). Such difference in the evolutionary strategies of thermophilic adaptation highlights long evolutionary distance between *T. maritima* and Archaea (9). Another hyperthermophilic bacteria *A. aeolicus* (21) also exhibits features typical for recolonization and development of sequence-specific strategy (see Table 3). At the same time, the composition-wise relationship between archaeal and bacterial parts of the proteome is not the same as in the *T. maritima* case (see Table 4). Slightly elevated packing density, compared with *T. maritima*, points to some role of structure-based stabilization (Table 2), which exists in *A. aeolicus* along with the sequence-based mechanism. This conclusion is consistent with the uniqueness of *A. aeolicus*'s evolutionary history, the deepest branched hyperthermophilic bacteria (21). Later events in protein evolution also affected sequences/structures of all species, bacterial and archaeal. For instance, contemporary *P. furiosus* features elevated content of charged residues compared with mesophilic *E. coli*, although not as pronounced as *A. aeolicus* or *T. maritima* (see Table 6). The diversity of the mechanisms of adaptation and ways underwent by different species leaves a room for further discussion of a role of recolonization and horizontal (lateral) versus vertical gene transfer (27), or even for challenging the very idea that life originated in a hot environment (28). However, we demonstrated here that adaptation can be generally considered from sequence- or structure-centric points of view. In particular, our findings and analysis highlight (*i*) physical mechanisms to achieve higher stability of a protein and (*ii*) the causal relationship between the physics of mechanisms of thermostability and adaptation strategies on the organismal level. Finally, a coherent viewpoint into the interplay of physical and evolutionary factors, provided by this analysis, can be potentially helpful in guiding our effort to design proteins with desired thermal properties.

We thank Jun Shimada for help with the unfolding simulations and William Chen for critical reading and valuable comments. I.N.B. is supported by the Merck Postdoctoral Fellowship for Genome-Related Research. This work was supported by National Institutes of Health Grant R01 52126.

1. Jaenicke, R. & Bohm, G. (1998) *Curr. Opin. Struct. Biol.* **8,** 738–748.
2. Berezovsky, I. N., Tumanyan, V. G. & Esipova, N. G. (1997) *FEBS Lett.* **418,** 43–46.
3. Schumann, J., Bohm, G., Schumacher, G., Rudolph, R. & Jaenicke, R. (1993) *Protein Sci.* **2,** 1612–1620.
4. Querol, E., Perez-Pons, J. A. & Mozo-Villarias, A. (1996) *Protein Eng.* **9,** 265–271.
5. Vetriani, C., Maeder, D. L., Tolliday, N., Yip, K. S., Stillman, T. J., Britton, K. L., Rice, D. W., Klump, H. H. & Robb, F. T. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 12300–12305.
6. Hurley, J. H., Baase, W. A. & Matthews, B. W. (1992) *J. Mol. Biol.* **224,** 1143–1159.
7. Thompson, M. J. & Eisenberg, D. (1999) *J. Mol. Biol.* **290,** 595–604.
8. Ogata, Y., Imai, E., Honda, H., Hatori, K. & Matsuno, K. (2000) *Origins Life Evol. Biosphere* **30,** 527–537.
9. Nelson, K. E., Clayton, R. A., Gill, S. R., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Nelson, W. C., Ketchum, K. A., *et al.* (1999) *Nature* **399,** 323–329.
10. Shimada, J., Kussell, E. L. & Shakhnovich, E. I. (2001) *J. Mol. Biol.* **308,** 79–95.
11. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) *J. Mol. Biol.* **247,** 536–540.
12. Stickle, D. F., Presta, L. G., Dill, K. A. & Rose, G. D. (1992) *J. Mol. Biol.* **226,** 1143–1159.
13. Corpet, F. (1988) *Nucleic Acids Res.* **16,** 10881–10890.
14. England, J. L. & Shakhnovich, E. I. (2003) *Phys. Rev. Lett.* **90,** 218101.
15. Go, N. & Abe, H. (1981) *Biopolymers* **20,** 991–1011.
16. Robic, S., Guzman-Casado, M., Sanchez-Ruiz, J. M. & Marqusee, S. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 11345–11349.
17. Hatanaka, H., Tanimura, R., Katoh, S. & Inagaki, F. (1997) *J. Mol. Biol.* **268,** 922–933.
18. Usher, K. C., de la Cruz, A. F., Dahlquist, F. W., Swanson, R. V., Simon, M. I. & Remington, S. J. (1998) *Protein Sci.* **7,** 403–412.
19. Dominy, B. N., Minoux, H. & Brooks, C. L., III (2004) *Proteins* **57,** 128–141.
20. Macedo-Ribeiro, S., Darimont, B., Sterner, R. & Huber, R. (1996) *Structure* **4,** 1291–1301.
21. Deckert, G., Warren, P. V., Gaasterland, T., Young, W. G., Lenox, A. L., Graham, D. E., Overbeek, R., Snead, M. A., Keller, M., Aujay, M., *et al.* (1998) *Nature* **392,** 353–358.
22. England, J. L., Shakhnovich, B. E. & Shakhnovich, E. I. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 8727–8731.
23. Shakhnovich, B. E., Deeds, E., Delisi, C. & Shakhnovich, E. (2005) *Genome Res.* **15,** 385–392.
24. Privalov, G. P. & Privalov, P. L. (2000) *Methods Enzymol.* **323,** 31–62.
25. Li, H., Helling, R., Tang, C. & Wingreen, N. (1996) *Science* **273,** 666–669.
26. Lawrence, J. G. (1999) *Curr. Opin. Microbiol.* **2,** 519–523.
27. Ochman, H., Lawrence, J. G. & Groisman, E. A. (2000) *Nature* **405,** 299–304.
28. Forterre, P. (1996) *Cell* **85,** 789–792.
29. Berezovsky, I. N., Namiot, V. A., Tumanyan, V. G. & Esipova, N. G. (1999) *J. Biomol. Struct. Dyn.* **17,** 133–155.
30. Berezovskii, I. N., Esipova, N. G. & Tumanian, V. G. (1998) *Biofizika* **43,** 392–402.